PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Cost-effectiveness of telehealthcare to patients with chronic obstructive pulmonary disease: Results from the Danish "TeleCare North" cluster-randomized trial.
AUTHORS	Witt Udsen, Flemming; Lilholt, Pernille; Hejlesen, Ole; Ehlers, Lars

VERSION 1 - REVIEW

REVIEWER	Padraig Dixon School of Social and Community Medicine, University of Bristol, UK
REVIEW RETURNED	27-Jan-2017

GENERAL COMMENTS	This paper describes a within-trial evaluation of a telehealth intervention implemented for COPD as part of a cluster RCT. My most substantial comment relates to the means by which the multiple imputation has been given effect – I suggest the next version of the manuscript reflects the clustered nature of the trial design and compares the results of so doing with the single level model that is described in the present version. 1.2 Specific comments p3 – Strengths and limitations, first bullet point – not clear what is meant by "requested by systematic reviews". Perhaps replace "requested" by "informed" Introduction p4 – It would be helpful to offer some brief (e.g. 2-3 sentences) explanation as to the nature of COPD, its impact on patients, which would help offer context for subsequent remarks made in relation to anxiety medication etc p5 – The acronym of GOLD in Table 1 should be explained and a reference provided (e.g. as done in the protocol paper) p6 – Description of medication use is unclear. As written, this suggests that patient-level medication use was not included in the analysis. However, Table 3 suggests that patient level data were collected. p6 – How were rehab costs, not collected due to IT implementation issues, imputed? These costs are noted as imputed in Table 4, but was this done as part of the overall multiple imputation, or by some other means? p6 – Pre-randomization costs were included to account for baseline differences in cost. Were such differences in cost expected, given the randomization procedures followed in the RCT? p7 – any particular reason for using a discount rate of 3%? p8 – It is interesting that SF-36 was collected in the trial as a means of estimating quality of life – what is the concordance between quality of life measured using this instrument and what is reported using EQ-5D? Could QALYs have been calculated from SF-36 and compared with those from EQ-5D, e.g. using the SF-6D algorithm?

p8 – there appears to be evidence of differential missingness – it would be useful to report a test of whether this is actually the case (it is likely to be the case give 325 in the telehealthcare group and 426 in the control group). Discussion of this aspect of the data should be offered, as it violates a “missing completely at random” assumption, although not necessarily a “missing at random” assumption.

p9 A further aspect of the imputation analysis that merits attention is the clustered nature of the RCT. The authors subsequently note the need to account for clustering in undertaking their analysis, but this consideration also applies to the imputation model used. It seems from the description that a single level imputation model has been used, with the clustering variable, which is not defined in the text, simply indicating allocation.

Gomes et al (http://journals.sagepub.com/doi/abs/10.1177/0272989X13492203) found that an MI approach that accounted for hierarchical/clustered nature of trial data performed better than a single level MI model. Diaz-Ordaz et al (http://onlinelibrary.wiley.com/doi/10.1111/rssa.12016/full) found that single level MI will tend to underestimate uncertainty. As an aside, the following paper on analyzing economic evaluations from cluster RCTs by Gomes et al is wort consideration https://www.ncbi.nlm.nih.gov/pubmed/22016450

I suggest that a multi-level imputation model be explored and its results compared to the single level model used

p9 – Do the authors have any concerns in using linear models for skewed cost and eq-5d distributions? This is not necessarily problematic, but it would be helpful to justify their use

p10 – Were sensitivity analyses pre-specified?

p11 – The number of non-smokers is lower in the intervention arm – was this accounted for (e.g. through minimisation) in the randomization?

p11 – The notes to Table 2 include reference to tests that do not seem to have been reported, e.g. Fisher’s (misspelled as Fischer in the text) exact test, Mann-Whitney

p13 – What specific impact does adjusting for baseline costs have on the between-arm comparisons? There is a very big difference between adjusted and unadjusted costs, the former of which reflects the influence of the named covariates in addition to baseline cost, and this raises the issue of why the adjustments seems to make such a large difference in the context of a randomized study design

p13 – There is a 10% difference in service cost before inclusion of intervention related costs – are there any possible explanations for this that may have influenced the results, e.g. participants in intervention arm were more aware of their condition and hence used more resources?

p14 – What mechanisms may have increased EQ-5D in the intervention arm?

p14 – How many individuals died in each arm, and what impact did deaths have on the QALY calculation?

p14 – What was the level of adherence to the intervention?

Discussion

p15 – I suggest removing the remark that the intervention might be cost-effective given the absence of an explicit Danish threshold value – this is speculative and the opposite conclusion also applies, and so a conservative interpretation of the results would refrain from making this conjecture. This remark also applies to the final sentence of the abstract

p16 – It is surprising to see speculation as to the stability of COPD borne by recruited patients - were these characteristics of patients

	not known from the RCT itself? p16 – A limitation of the analysis is that separate models were estimated for costs and effect, so probably less efficient than joint modelling of both using correlated covariance structures p16 – Not clear what this sentence means “However, this does not exclude that the COPD subgroup is cost-effective which remain to be seen.” Is a subgroup analysis of the WSD project planned in relation to COPD patients? p17 – What is the “optimal implementation” referred to in the 2nd paragraph? p17 – A synthesis of the evidence in relation to telehealth for chronic conditions is available in (https://www.journalslibrary.nihr.ac.uk/pgfar/pgfar05010/#/abstract)

REVIEWER	Wang Wenru National University of Singapore Singapore
REVIEW RETURNED	06-Feb-2017

GENERAL COMMENTS	Thanks for the opportunity to review this paper. The paper addressed an important research issue of cost-effectiveness of a telehealthcare intervention. It is generally well-written, yet a few grammatical errors noted e.g. page 2, line 11: it should be ‘A 12-month cost-utility analysis…’. The authors may double check the grammar when resubmit their paper. Nevertheless, I have only minor comments for authors to consider. 1. The conclusion of the study is ambiguous. The authors stated that ‘Telehealtcare is unlikely to be cost-effective addition to usual care if it is offered to all patients with COPD’. However, the authors also concluded that ‘Since no willingness-to-pay threshold exists in Denmark, it may still be cost-effective here”. Such conclusion may confuse the readers, especially without knowledge on what ‘willingness-to-pay threshold’ refers?? A clearer conclusion from the study is needed. 2. Table 1 is too busy and not reader-friendly. Table 1 may not need. I understand the authors have published their study protocol, and details of the study participants, design and interventions have been presented in the published protocol. The authors may simply describe the study methods in terms of design, participants, and interventions and cite their published protocol paper. 3. The research ethics (e.g. participant consent, ethics approval) need to be addressed in the method section appropriately. 4. You may revise the subheading of ‘Effectiveness” on page 8. What does this refer? Would be good to make it more clear and specific. 5. I cannot comment too much on cost-effectiveness analysis and sensitivity analysis as these are beyond my expertise. A review done by a biomedical statistician may be needed to sure the authors used the right statistical methods. 6. Results: Tables may need to revised based on journal’s format. For table 2, we don’t normally report raw differences, instead, the p values indicating the significant difference between two groups would be reported. 7. Discussion should start and focus on what you have found from your study and comparison with other studies. A brief conclusion of the study should be included.

VERSION 1 – AUTHOR RESPONSE

We would like to thank both reviewers for extremely useful comments that we feel have helped improve our paper significantly. We have inserted our response point by point in the following.

p3 – Strengths and limitations, first bullet point – not clear what is meant by “requested by systematic reviews”. Perhaps replace “requested” by “informed”
We agree. Thank you for the suggestion.
The formulation has been changed.

p4 – It would be helpful to offer some brief (e.g. 2-3 sentences) explanation as to the nature of COPD, its impact on patients, which would help offer context for subsequent remarks made in relation to anxiety medication etc
We thank the reviewer for the suggestion.
2 sentences have been added to the introduction describing the definition and nature of COPD.

p5 – The acronym of GOLD in Table 1 should be explained and a reference provided (e.g. as done in the protocol paper)
We agree.
GOLD has been explained and a reference provided as in the trial protocol in Table 1.

p6 – Description of medication use is unclear. As written, this suggests that patient-level medication use was not included in the analysis. However, Table 3 suggests that patient level data were collected.
We agree.
The formulation has been changed to make it clearer that patient-level medication was collected.

p6 – How were rehab costs, not collected due to IT implementation issues, imputed? These costs are noted as imputed in Table 4, but was this done as part of the overall multiple imputation, or by some other means?
It was done as part of the overall multiple imputation.
This is stated on p.9 in the paragraph “Missing data”.

p6 – Pre-randomization costs were included to account for baseline differences in cost. Were such differences in cost expected, given the randomization procedures followed in the RCT?
We suspected that it could be the case.
Healthcare delivery is decentralized in Denmark and all sectors and institutions, such as the municipality districts have a high degree of autonomy in visitation of service use to its citizens. This implicate that there is a risk that variations could occur that are attributed to differences in practices as opposed to differences in health or socio-demographic status when the unit of randomization is municipality districts. This argument has been added to the manuscript.

p7 – any particular reason for using a discount rate of 3%?
To our knowledge, discount rates usually vary between 3-5% in applied health economic research http://onlinelibrary.wiley.com/doi/10.1111/j.1524-4733.2004.74002.x/full. In national Danish national capital accounting, a discount rate of 3% for IT-equipment may be used and a reference has been provided. As we understand it, it is slightly lower than in the UK, where a discount rate of 3.5 is suggested, so the difference is not big.

p8 – It is interesting that SF-36 was collected in the trial as a means of estimating quality of life – what is the concordance between quality of life measured using this instrument and what is reported using EQ-5D? Could QALYs have been calculated from SF-36 and compared with those from EQ-5D, e.g. using the SF-6D algorithm?

The scope of this paper is to report the within trial cost-effectiveness results and SF-36 is not often used in cost-effectiveness research when EQ5D data is available also.
The SF-36 has been used to report the effectiveness of the trial in another paper.
Collecting data on two generic instruments would allow for a later comparison of concordance between the instruments. And incidentally, both instruments will also be used in another clinical trial of telehealth for patients with chronic heart failure. So, a methodology paper could be a future option for us.

p8 – there appears to be evidence of differential missingness – it would be useful to report a test of whether this is actually the case (it is likely to be the case give 325 in the telehealthcare group and 426 in the control group). Discussion of this aspect of the data should be offered, as it violates a “missing completely at random” assumption, although not necessarily a “missing at random” assumption.
We agree. One could almost say that if data was MCAR, it would be a surprise.
The association between missingness of outcome variables and all collected baseline variables were tested by a series of logistic regressions. Some baseline variables proved statistically significant ($p<0.05$), so data are not missing completely at random (MCAR). Missingness of the EQ5D summary score was also associated with the EQ5D summary score at baseline.

An assumption that data are missing at random (MAR) is therefore inserted and can to our knowledge not be testet. MAR can be a plausible assumption if a wide range of variables, and variables that are predictive of missingness, are included in the imputation model. Also, more detail on the multiple imputation procedure has been added to the manuscript in the section containing “Missing data”.

p9 A further aspect of the imputation analysis that merits attention is the clustered nature of the RCT. The authors subsequently note the need to account for clustering in undertaking their analysis, but this consideration also applies to the imputation model used. It seems from the description that a single level imputation model has been used, with the clustering variable, which is not defined in the text, simply indicating allocation.
Gomes et al (http://journals.sagepub.com/doi/abs/10.1177/0272989X13492203) found that an MI approach that accounted for hierarchical/clustered nature of trial data performed better than a single level MI model. Diaz-Ordaz et al (http://onlinelibrary.wiley.com/doi/10.1111/rssa.12016/full) found that single level MI will tend to underestimate uncertainty. As an aside, the following paper on analyzing economic evaluations from cluster RCTs by Gomes et al is wort consideration
https://www.ncbi.nlm.nih.gov/pubmed/22016450
I suggest that a multi-level imputation model be explored and its results compared to the single level model used
We thank the reviewer for drawing out attention to multi-level multiple imputation.

We can see that the reviewer’s university (the University Bristol) has done a lot of work on multilevel modelling in collaboration with London School of Hygiene & Tropical Medicine including multi-level multiple imputation and are at the forefront of publishing methodology papers on the subject and software to be utilized.

As practitioners of cost-effectiveness analysis we have not quite caught up with this development primarily because multi-level multiple imputation is (still) not part of conventional statistical software. Although specialized software do exist (e.g. REALCOM impute developed by the schools mentioned above), we are not capable of applying these programs in a foreseeable future, although we would certainly explore opportunities of conducting multi-level imputation within these programs from now on.
On the other hand (and this should not be an excuse, just to put our approach into perspective), we have used multiple imputation with cluster as a fixed effect, which is more than most studies that were

found in a review by Fiero et. al. published in 2016 that investigated how missing data was handled in 86 published cluster-randomized trials (https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-016-1201-z). 55% were complete case analyses and 8% single imputation (mean or LOCF imputation). Furthermore, we have been struggling to find other empirical cost-effectiveness studies that use multi-level multiple imputation, although some cost-effectiveness study protocols have been published recently with the intent of conducting multi-level multiple imputation.

We have inserted the following in the Discussion as a first limitation of our study:
"A limitation of the study is that single-level multiple imputation with clustering as a fixed effect was performed. Gomes et. al. has found that an imputation approach that account for clustering as a random effect perform better than single-level imputation (http://journals.sagepub.com/doi/abs/10.1177/0272989X13492203). Furthermore, Andridge have in a simulation study found that including clustering as a fixed effect in the imputation model could overestimate the uncertainty of the estimates, especially if the number of clusters are small and the ICC is low as in this case (https://www.ncbi.nlm.nih.gov/pubmed/21259309). However, a barrier to the adoption of multi-level multiple imputation is that these techniques are not part of conventional statistical software".

In addition, the degree of missingness in this study has been highlighted in the "strengths and limitations" bullet points beneath the Abstract.
We are hoping that this would be satisfactory.

p9 – Do the authors have any concerns in using linear models for skewed cost and eq-5d distributions? This is not necessarily problematic, but it would be helpful to justify their use
It is always difficult to choose analysis strategy as described by Nixon et. al. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3470917 – a decision that can be even more difficult for cost-effectiveness analysis of cluster-randomized trials, because this field is very much in development meaning that analyses that simultaneously can account for clustering, skewed outcomes/costs and correlation of costs/outcomes are not part of conventional statistical software packages yet, making it difficult for practitioner of cost-effectiveness analyses to utilize them.
This means that none of the standardly available techniques is usually optimal given the characteristics and behaviour of your own data. Had we chosen traditional SUR-models (to account for correlation between outcomes/costs) or GLM-models (to account for skewed data), we would be open to other criticism (not accounting for clustering).

We decided to use simple multilevel models with normally distributed errors terms, which are also described in the trial protocol. This came as a result of the work of Bachmann and colleagues that have recommended the xtmixed procedure as one way of analyzing cost-effectiveness data (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2020454/). Gomes et. al. that the reviewer has suggested for us, have incidentally also demonstrated that normally distributed multilevel models perform well under the conditions of our study (unequal cluster sizes and a small number of clusters. See https://www.ncbi.nlm.nih.gov/pubmed/22016450: Although Gomes uses an MLM with normally distributed error terms that accounts for correlation of outcomes and costs (a bivariate multilevel model) in the basecase analysis (0.2 individual level and 0 cluster-level), the multiway sensitivity analysis combined scenarios with different levels of correlations in outcomes and costs (ranging from -0.5 to 0.5) and concludes: "For the MLM and TSB (with shrinkage correction), the CI coverage remains relatively good even when the study has few highly imbalanced clusters and highly skewed costs. In further scenarios that combined variation in cluster-size imbalance and number of clusters with other parameters, such as different levels of individual- and cluster-level correlation, all methods performed well except in scenarios with few clusters, where SUR and GEE reported poor coverage" (p.356). We have interpreted this to mean that a MLM scenario where this correlation was 0, or at least small, was simulated with similar results.

We have been unsure of how much detail and discussion that is requested and have therefore checked the reviewer’s own arguments for choice of analysis strategy in the Healthlines studies and inserted that we rely on near-normality in the Methods section. In addition, a discussion describing the arguments from above has also been added to the manuscript in the Discussion and we are hoping that this is satisfactory.

p10 – Were sensitivity analyses pre-specified?
Good question. No they were not.
They are meant to reflect the real decision-uncertainty that decision-makers had that followed the presentation of the base case analysis. This is made clearer in the manuscript.

p11 – The number of non-smokers is lower in the intervention arm – was this accounted for (e.g. through minimisation) in the randomization?
Good question. The question refers to how imbalance in a prognostic factor could influence the outcomes of the trial.

No minimization/stratification was used for patients in the randomization procedure. With what we know now, a minimization process might have been desirable because the treatment effect (QALY) is so small. But the large sample size meant that we did not think of it in the design of the trial which took place in 2011-2012 before higher quality evidence of cost-effectiveness were published that demonstrated small incremental QALYs. Regardless, minimization/stratification would probably be a good point to remember for future research.

If an inference test is used for smoking status at baseline, e.g. a Fisher’s exact test, the difference in smoking status between intervention and control group is not statistically significant (p-value=0.103) and including smoking status as an additional covariate in the QALY model have little effect on incremental QALYs (from 0.01316 without to 0.01288 with smoking status included). The same is true when including smoking status in the totalcost model, where incremental costs falls to €705 with smoking status included as an additional covariate (from €728 without its inclusion).

A statement of the higher proportion of smokers in the intervention arm is added to the Results and a discussion reflecting the lack of minimization is provided as a limitation in the Discussion.

p11 – The notes to Table 2 include reference to tests that do not seem to have been reported, e.g. Fisher’s (misspelled as Fischer in the text) exact test, Mann-Whitney
Thank you.
We have used a template and forgot to change it to the reporting of the results and the tests conducted. The sentences have been deleted.

p13 – What specific impact does adjusting for baseline costs have on the between-arm comparisons? There is a very big difference between adjusted and unadjusted costs, the former of which reflects the influence of the named covariates in addition to baseline cost, and this raises the issue of why the adjustments seems to make such a large difference in the context of a randomized study design
We thank the reviewer for making this point.

In the Discussion the following is added.
“It was suspected that baseline differences in costs could occur that would not necessarily be explained by differences in health or socio-demographic characteristics, e.g. due to variations in visitation practice across municipality districts. If baseline cost is removed as a covariate in the analysis of adjusted total costs, incremental costs rise from €728 to €1334.

Recent guidance from the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) suggest that baseline resource use should be collected and that the analysis of both costs and effects could include baseline measures of costs https://www.ispor.org/Cost-Effectiveness-clinical-trials-guideline.pdf (p. 164 and p 167) which is also suggested by van Asselt et. al. https://www.ncbi.nlm.nih.gov/pubmed/19640014. However, guidance is not as explicit as including baseline utility in the analysis of QALYs https://www.ncbi.nlm.nih.gov/pubmed/15497198 and https://www.ispor.org/Cost-Effectiveness-clinical-trials-guideline.pdf. In our opinion, the baseline difference in cost reported in this study underlines the importance of requesting information on institutional context, such as variations in existing resource patterns, when interpreting cost-effectiveness research".

We hope that this is satisfactory.

p13 – There is a 10% difference in service cost before inclusion of intervention related costs – are there any possible explanations for this that may have influenced the results, e.g. participants in intervention arm were more aware of their condition and hence used more resources?
Thank you for the suggestion.
The rationale is sound and this argument would be much more in focus in another manuscript tapping into potential sources of heterogeneity in cost-effectiveness. Stating much more than what we have added in the Implications paragraph of the Discussion is therefore premature:

"There was a 10% difference in service cost before inclusion of intervention related costs and possible explanations could be that patients randomized to telehealthcare became more aware of their disease and hence used more resources or it could be that especially municipalities discovered patients with an unmet need for home care when telehealthcare was introduced. Future research planned within this trial would seek to tap into possible explanations for this difference".

p14 – What mechanisms may have increased EQ-5D in the intervention arm?
Good question. We have not investigated if "fits", "visibility" or "relationships" could be mechanisms https://implementationscience.biomedcentral.com/articles/10.1186/s13012-015-0238-9, nor have we identified our own mechanisms in the design of our study, if that is what the reviewer is suggesting. But mechanisms might be a further topic for our planned heterogeneity analysis.

Our study design was not theoretically grounded such as the Healthlines studies that the reviewer has been part of. We have therefore treated intervention-mechanisms as black boxes. We agree that future studies of complex telehealth interventions should have a more theoretical starting point. We have therefore ended the Discussion with the following:

"Telehealthcare is a complex intervention involving not only a broad class of technologies, but also organizational infrastructures, actions of healthcare professionals and patients. Experimental evaluation research has been criticized for being a-theoretical in nature in trying to understand why and under what circumstances complex interventions are (un)likely to lead to desired outcomes, (Pawson and Tilley). In this study, mechanisms leading to higher health-related quality of life and cost in the telehealthcare group has largely been treated as a black-box, where patient education, monitoring, emotional support, assisted planning etc. could all have an effect (McLean). We would recommend, that future cost-effectiveness studies are more informed by a program theory, such as the TECH model (Salisbury) that were used in the Healthlines studies (Dixon1 and Dixon2). These studies sought to describe contexts or account for the causation of the most important telehealthcare-activities that were most likely to activate mechanisms that could lead to "efficient" design and deployment of telehealthcare as well as the context in which telehealthcare is implemented.

We hope that this would be satisfactory.

p14 – How many individuals died in each arm, and what impact did deaths have on the QALY calculation?
The number of dead patients in each arm is now provided p9.

We are not sure what the reviewer means by the impact on the QALY calculation. Is the reviewer specifically interested in how deceased patients were scored at follow-up (information now provided p9) or are the reviewer suggesting some form of analysis - e.g. by not providing deceased an EQ5D summary score of 0 and then do the QALY analysis again in order to test the robustness of the results to death, when incremental QALYs are so small?

We have therefore inserted this paragraph in the Discussion, p18:
"When interpreting small differences in effectiveness, it is important to be aware that results can be highly sensitive to between-group differences in death. Even though, it is standard practice to assign an EQ5D summary score of 0 to deceased patients (41) in order to calculate incremental QALYs, this practice could potentially have a drastic effect on estimated cost-effectiveness. However, in this case the estimated between-arm QALY difference from the imputed dataset and an analysis where this EQ5D scoring is not done, are similar (QALY difference reduced from 0.01316 to 0.01004)".

p14 – What was the level of adherence to the intervention?
We thank the reviewer for drawing our attention to adherence.
Data on each monitoring contact was available for 21 of the 26 municipality districts included (the remaining 5 districts has reported aggregated time spent monitoring each participant during the trial-period). The median number of contacts was 53 out of a total of 64 planned (daily first two weeks and once a week in the following 50 weeks). Although compliance to monitoring is not the sole source of adherence to the intervention and we do not have complete data for each individual encounter, it does suggest a high commitment to the TeleCare North initiative.
This has been added to the Discussion.

p15 – I suggest removing the remark that the intervention might be cost-effective given the absence of an explicit Danish threshold value – this is speculative and the opposite conclusion also applies, and so a conservative interpretation of the results would refrain from making this conjecture. This remark also applies to the final sentence of the abstract
We agree.
The ambiguity in the formulation has been deleted on p.15 as well as in the final sentence of the abstract.
p16 – It is surprising to see speculation as to the stability of COPD borne by recruited patients - were these characteristics of patients not known from the RCT itself?
OK. These characteristics were known.
We have just wondered if conclusions would have been different if we had recruited patients from open hospital admissions instead of relying on general practitioners to refer patients to the telehealth intervention.

The formulation has been deleted.

p16 – A limitation of the analysis is that separate models were estimated for costs and effect, so probably less efficient than joint modelling of both using correlated covariance structures
We agree.
We have added a paragraph stating this to the Discussion as a limitation

p16 – Not clear what this sentence means "However, this does not exclude that the COPD subgroup is cost-effective which remain to be seen." Is a subgroup analysis of the WSD project planned in

relation to COPD patients?
Thank you.
The last part of the sentence ("which remain to be seen") has been deleted, since we do not know if further subgroup analyses will be reported.

p17 – What is the "optimal implementation" referred to in the 2nd paragraph?
Thank you.
The reviewer is right in suggesting that there is probably no implementation that is "optimal". The intention was to explain that cost-effectiveness might be improved by focusing on improving implementation in the future by learning from the experiences gained in the study. However, this is also stated in the sentence that follows. The sentence has therefore been deleted.

Reviewer 2: It is generally well-written, yet a few grammatical errors noted e.g. page 2, line 11: it should be 'A 12-month cost-utility analysis…'. The authors may double check the grammar when resubmit their paper. Nevertheless, I have only minor comments for authors to consider.
Thank you. And we agree that proof editing should be done.
The "s" in months has been removed and could have been an oversight. A proof-editor has been used prior to submission that we have used several times to comment on the manuscript, so we are unsure how to proceed.

The conclusion of the study is ambiguous. The authors stated that 'Telehealtcare is unlikely to be cost-effective addition to usual care if it is offered to all patients with COPD'. However, the authors also concluded that 'Since no willingness-to-pay threshold exists in Denmark, it may still be cost-effective here". Such conclusion may confuse the readers, especially without knowledge on what 'willingness-to-pay threshold' refers?? A clearer conclusion from the study is needed.
We agree. Thank you.
The ambiguity in the formulation has been deleted on p.15 as well as in the final sentence of the abstract.

Table 1 is too busy and not reader-friendly. Table 1 may not need. I understand the authors have published their study protocol, and details of the study participants, design and interventions have been presented in the published protocol. The authors may simply describe the study methods in terms of design, participants, and interventions and cite their published protocol paper.
Good point.
We have been unsure ourselves of how much detail is requested, when we have published a study protocol. We have therefore used Table 1 as a form of compromise between those that prefer a lot of details, so that the manuscript can be read independently of the protocol and those who do not.

We have previously submitted this manuscript to the BMJ, where 2 reviewers (as with reviewer 1 in this review) did not comment on the readability of Table 1.

If an editorial decision is made to follow this reviewer's suggestion, we would of course remove Table 1 and include a short description in the body of text.

The research ethics (e.g. participant consent, ethics approval) need to be addressed in the method section appropriately.
We are not sure what is meant by the comment?
According to the current CONSORT checklists and the instructions for authors from BMJOpen, we can find no requirement for including research ethics in the Methods section. An ethics statement is made at the end of the manuscript, where they are usually placed. Does the reviewer mean a description of how informed written consent was obtained? This is described in the study protocol p.3 and typically not restated in an economic evaluation.

You may revise the subheading of 'Effectiveness" on page 8. What does this refer? Would be good to make it more clear and specific.
We are unsure what is meant by this comment and think that this comment might be due to unfamiliarity with cost-effectiveness evaluation?
In a cost-effectiveness study it is normal to have two paragraphs in the Methods section describing measurement and valuation of effectiveness and measurement and valuation of resource use.
We have tried to compare our effectiveness description (content, order and detail) with other research on cost-effectiveness for telehealth, e.g. the Whole System Demonstrator study (http://www.bmj.com/content/346/bmj.f1035) and the Healthlines studies (http://bjpo.rcpsych.org/content/2/4/262 and http://bmjopen.bmj.com/content/6/8/e012352.full). We find only a few minor differences in formulations, but content, order and detail are very similar.

I cannot comment too much on cost-effectiveness analysis and sensitivity analysis as these are beyond my expertise. A review done by a biomedical statistician may be needed to sure the authors used the right statistical methods.
We agree.
And Reviewer 1 seems to be highly qualified.

Results: Tables may need to revised based on journal's format. For table 2, we don't normally report raw differences, instead, the p values indicating the significant difference between two groups would be reported.
We disagree.
This is a common misunderstanding for randomized trials. The reviewer's suggestion implicate that we test for the probability that observed baseline differences could have occurred by chance. However, we already know that any differences are caused by chance (due to randomization), so the test does not make sense.
According to the current CONSORT checklist item 15 for reporting baseline information it is stated that "the variability of the data should be reported, along with average values. Continuous variables can be summarized for each group by the mean and standard deviation" Categorical variables should be presented as numbers and proportions for each category. And this is what we have done.
In fact, it is directly stated that inferential statistics should not be used: "Standard errors and confidence intervals are not appropriate for describing variability—they are inferential rather than descriptive statistics".
See http://www.consort-statement.org/checklists/view/32-consort/510-baseline-data
Calculating the raw difference is meant as a service to the reader, so they do not have to do it themselves.

Discussion should start and focus on what you have found from your study and comparison with other studies. A brief conclusion of the study should be included.
We agree, thank you.
Although we were stating our primary findings first by presenting our ICER (the main cost-effective result) and describing how sensitive this result is for changed assumptions (the results from our sensitivity analyses), we could also have given more attention to the adjusted main differences in QALYs and costs. These are not results that one necessarily would highlight in a full-scale economic evaluation as ours (see Drummond, M., Sculpher, M., Claxton, K., Stoddart, G., & Torrance, G. (2015). Methods for the Economic Evaluation of Health Care Programmes (Fourth edi). Oxford University Press). Regardless, these have been added to the Discussion.

VERSION 2 – REVIEW

REVIEWER	Padraig Dixon University of Bristol, UK

REVIEW RETURNED	17-Mar-2017

GENERAL COMMENTS	I am grateful to the authors for their considered response to the various comments and suggestions made in the first round of comments. One final observation, which the authors may or may not wish to include: Mechanisms of effect: It may be worth noting in the main text of the paper that even where a theory-driven programme has been used to guide the development of a telehealth intervention, the mechanisms that gave rise to between-arm differences in EQ-5D were still difficult to identify, reflecting the fact that trials are generally designed to assess outcomes other than EQ-5D, and it is the mechanisms underlying these other outcomes (e.g. encouraging smoking cessation or weight loss) that tend to receive more attention in program development

REVIEWER	Wang Wenru National University of Singapore Singapore
REVIEW RETURNED	19-Mar-2017

GENERAL COMMENTS	Thanks for authors' effort in addressing the comments and revsing the paper. Although some of my comments were not fully addressed in the revised paper, the authors made clear responses. I have no furhter comments and all the best.

VERSION 2 – AUTHOR RESPONSE

Reviewer 1:
One final observation, which the authors may or may not wish to include: Mechanisms of effect: It may be worth noting in the main text of the paper that even where a theory-driven programme has been used to guide the development of a telehealth intervention, the mechanisms that gave rise to between-arm differences in EQ-5D were still difficult to identify, reflecting the fact that trials are generally designed to assess outcomes other than EQ-5D, and it is the mechanisms underlying these other outcomes (e.g. encouraging smoking cessation or weight loss) that tend to receive more attention in program development.

Response:
We thank the reviewer for the suggestion. We have inserted the following to conclude the Discussion: “However, context and mechanisms that specifically gave rise to between-arm differences in EQ-5D in the Healthlines studies are difficult to identify, reflecting that program theories are often focused on explaining trial-related aspects or outcomes (e.g. smoking cessation or weight loss). In the future, context and mechanisms leading to between-arm differences in EQ-5D and costs should receive more attention in program theory development.”